Average time until fixation of a mutant allele in a given population

KOMI MESSAN

Abstract

One of the important problems in population genetics is how long it takes for a gene to go to fixation (become established). A mutant gene in a given population will eventually be lost or established. The particular interest of this research is to know the mean time for a mutant gene to become fixed in a population, and we will exclude the case when this gene is lost. A diploid population of N individuals will be considered with a forward and backward mutation of u and v respectively per basis. Using a set of nonlinear equations, we will first calculate the genotype frequencies which will allow us to find the equilibrium points for the infinite population. With the diffusion theory, we will approximate the time to fixation for finite populations. We will then proceed with a numerical approximation using C++ to see a close result for the problem.

1 Introduction

The main idea in population genetics is evolution. Evolution is much different from most studies in biology for the fact that its insights are theoretical rather than observational or experimental. Most evolutionary studies concern the frequencies or the fitnesses of genotype in a given population. Evolution is the change in the frequencies of genotype through time, perhaps due to their differences in fitness (Gillespie 2004). Evolution can also be explained by two forces: forces that introduce variation in phenotypic character such as eye colors, height or certain behaviors and forces that make some traits become rare or more common. The main cause of variation is mutation, which changes the sequence of a gene (Strickberger 2000). In other words, mutation is a change in the DNA sequence of Cell's genome.

The forces that make traits to become common or rare are caused by two main processes. One of these processes is natural selection which is a key term used in genetic evolution (Strickberger 2000). Natural selection is the differential reproductive success of a any given organism. Very often, organisms produce more offspring than their environment can support; because of this, not every individual in a population survives in the generation and this can be one of the main cause of natural selection. Over many generations mutation produces random changes in traits, which are then filtered by natural selection and the beneficial traits retained (Gillespie 2004). Another cause of evolution is genetic drift, which is a change in the relative frequencies in which gene variant occurs

in a population due to random sampling and chance. These random changes affect evolution in two important ways. First, a dispersive force that removes genetic variations from population. Let us note that the rate of removal is very weak since it is inversely proportional to the population size. Tahe other is the effect of drift on the probability of survival of a new mutation (Gillespie 2004).

Another evolutionary process we will see later in this paper is genetic recombination. In this process, the DNA or sometimes the RNA molecule breaks and then joins another DNA molecule. Recombination can also have a big impact on the evolutionary processes and this was shown by Paul G. Higgs (Higgs 1997). We will take the same approach to show this but in a higher dimension.

2 Background

We will consider a diploid population consisting of N individuals and having the variance effective number N_e. Let us note that N_e may be different than N and a good explanation of N_e can be found in "KIMURA and CROW 1963". Throughout this paper, we will develop a model that has been introduced by Paul G. Higgs (Higgs 1997) and some other authors (Michalakis and Slatkin, 1996; Phillips, 1996; Stephan, 1996). Most of these authors develop a model in which mutation is irreversible but we will consider a reversible mutation in this paper using the same model.

Our model will involve 2 loci, each with two alleles. The two alleles will be labeled A and a at one locus and B and b at the other. We will therefore have four genotypes: ab, Ab, aB, and AB. The frequencies of ab and AB will respectively be denoted x_0 and x_2. Both the double mutant genotypes have a frequency denoted x_1. We are therefore assuming these double mutant genotype have the same fitness.T genotype ab has fitness 1 and AB has fitness $1 - s_2$. However, the double mutant genotypes have fitness $1 - s_1$. Let us note that $x_0 + 2x_1 + x_2 = 1$.

Throughout this paper, we will assume both u and v are $< 10^{-6}$, and both s_1 and s_2 will be in the range 0.01 to 0.005. For different values of selection, mutations (forward and backward mutations), computer simulation will be used to approximate the time at which the first allele will arrive at genotype AB. Consider it starts from the genotype ab.

3 Equilibrium points of the infinite population

Prior to looking at the changes in the finite population, we will first look at the genotype frequencies in the infinite population. If we call $x_0, x_1, and x_2$ the frequencies at generation t then the frequencies at generation $t + 1$ will be denoted $X_0^*, X_1^*, and X_2^*$. Considering all our parameters are different from zero, we obtain the following set of nonlinear equations:

$$(X_0)^* = (1 - 2u)x_0 + 2vx_1 + 2s_1x_0x_1 - r\left(-x_1^2 + x_0x_2\right) \tag{1}$$

$$(X_1)^* = 2ux_0 + (2 - 2u - 2v)x_1 - 2s_1x_0x_1 + 2vx_2 + 2(-s_1 + s_2)x_1x_2 + 2r\left(-x_1^2 + x_0x_2\right) \quad (2)$$

$$(X_2)^* = 2ux_1 + (1 - 2v)x_2 - 2(-s_1 + s_2)x_1x_2 - r\left(-x_1^2 + x_0x_2\right) \quad (3)$$

Because of the complexity of the equations, we will try to simplify the equations by setting recombination to be zero. Doing so, we get these following set of equations:

$$(X_0)^* = (1 - 2u)x_0 + 2vx_1 + 2s_1x_0x_1 \quad (4)$$

$$(X_1)^* = 2ux_0 + (2 - 2u - 2v)x_1 - 2s_1x_0x_1 + 2vx_2 + 2(-s_1 + s_2)x_1x_2 \quad (5)$$

$$(X_2)^* = 2ux_1 + (1 - 2v)x_2 - 2(-s_1 + s_2)x_1x_2 \quad (6)$$

At the fixed position, $X_0^* = x_0, X_1^* = x_1, and X_2^* = x_2$ (Higgs 1997). Hence solving for all the three variables $(x_0, x_1, and x_2)$, we get the following equilibrium frequencies:

$$x_0 = -\frac{v}{s_1} - \frac{2uv}{{s_1}^2} \quad (7)$$

$$x_1 = \frac{1}{2} - \frac{u}{2s_2} + \frac{v}{2s_1} + \frac{uv}{{s_1}^2} + \frac{5{s_1}^4uv}{{s_2}^6} + \frac{-\frac{s_1u}{2} + uv}{{s_2}^2} + \frac{-\frac{{s_1}^2u}{2} + 2s_1uv}{{s_2}^3} + \frac{-\frac{{s_1}^3u}{2} + 3{s_1}^2uv}{{s_2}^4} + \frac{-\frac{{s_1}^4u}{2} + 4{s_1}^3uv}{{s_2}^5} \quad (8)$$

$$x_2 = -\frac{37{s_1}^4}{{s_2}^4} + \frac{7{s_1}^3}{{s_2}^3} - \frac{{s_1}^2}{{s_2}^2} + \frac{s_1}{s_2} \quad (9)$$

Now because we assumed $u^2 = 0$ and $v^2 = 0$ then we can say that $u^2 \approx v^2 \approx uv$. The equations (7), (8), and (9) hence become

$$x_{0=} - \frac{v}{s_1} \quad (10)$$

$$x_1 = \frac{1}{2} + \frac{v}{2s_1} - \frac{us_1^4}{2s_2^5} - \frac{us_1^3}{2s_2^4} - \frac{us_1^2}{2s_2^3} - \frac{us_1}{2s_2^2} - \frac{u}{2s_2} \quad (11)$$

$$x_{2=} - \frac{37{s_1}^4}{{s_2}^4} + \frac{7{s_1}^3}{{s_2}^3} - \frac{{s_1}^2}{{s_2}^2} + \frac{s_1}{s_2} \quad (12)$$

We can clearly see that equations (9) and (12) are the same. This is because the frequency x_2 does not depend on any mutation rate after simplification of the original solutions. In the above solutions, Both u and v are less than s_1 and s_2. These are the simplified version of the original solutions. We also obtain a number of complex solutions, but we are only interested in the real solution as written above. Here we assumed that all mutations (u and v) to the power ≥ 2 are equal to zero since since $u < 10^{-6}$ and $v < 10^{-6}$. Also because the selection coefficients are very low (between 0.01 and 0.005), we assumed $s_1^n = 0$ and $s_2^n = 0$ whenever n ≥ 5.

4 Dynamics of finite populations

Now we will look into the change in the finite population. Let's remember there are four genotype and their frequencies must be equal to 1 ($x_0 + 2x_1 + x_2 = 1$). Hence there are four independent frequency variables, but we are assuming the two single mutants (Ab and aB) are the same and have the same fitnesses of $1 - s_1$. This assumption leads us to work with a three dimensional system. Since we know that the total frequency is 1, it will be easier to work only with two variables and once we get the results, we can find the third variable in term of the others. Our system is therefore reduce to a two dimensional system. Previously Kimura and Ohta (1968) have developed a 1 dimensional model using the diffusion models. Higgs (1997) has also shown that it is possible to solve a 1D system with the diffusion models. We will also use the same model to solve our problem.

Here, because of the mutation and selection forces, we need a drift term which we will call $m(x)$. The $m(x)$ or the infinitesimal mean is the change of frequency in one generation. $m(x)$ can also be called the expected mean change in our variable of interest. A variance and covariance will be needed since they follow a multinomial distribution. Our variance and covariance will respectively be denoted $v(x_i, x_i)$ and $cov(x_i, x_j)$ since we are working in 2 dimensional system. From Lynch's appendix (2008), we see that

$$v(x_i, x_i) = \frac{x_i * (1 - x_i)}{2 * N_e} \tag{13}$$

$$cov(x_i, x_j) = \frac{x_i * x_j}{2 * N_e} \tag{14}$$

We could use the Kolmogorov forward equation (or KFE) as described by Kimura and Ohta (1968), but since our system is two dimensional this diffusion model will not work for us. We will instead use the extended KFE shown by Lynch in his appendix (2008)

$$\frac{\partial[\rho(x,p,t)]}{\partial t} = \frac{1}{2}\sum_{i=1}^{k-1}\frac{\partial^2}{\partial x_i^2}[\rho(x,p,t)\frac{x_i(1-x_i)}{N_e}]$$
$$-\sum_{i<j}\frac{\partial^2}{\partial x_i \partial x_j}[\rho(x,p,t)\frac{x_i x_j}{N_e}] - \sum_{i=1}^{k-1}\frac{\partial m(x)\rho(x,p,t)}{\partial x_i} \quad (15)$$

where the first part of the equation (15) is the allele-frequency variances, the second part involves the covariances between allele frequencies, and the third involves the change of frequency in generation or the mean. In this equation (15) $\rho(x,p,t)$ denotes the density function with x being the vector of allele frequencies, p the vector of their starting values, and t the time (Lynch appendix 2008). Applying this extended KFE (15) to our specific model, we get

$$\frac{\partial[\rho(x_0,x_2,t)]}{\partial t} = [\frac{1}{2}\frac{\partial^2}{\partial^2 x_0}(\rho(x,p,t)\frac{x_0(1-x_0)}{2N_e}) + \frac{1}{2}\frac{\partial^2}{\partial^2 x_2}(\rho(x,p,t)\frac{x_2(1-x_2)}{2N_e})] - [\frac{\partial^2}{\partial x_0 \partial x_2}(\rho(x,p,t)\frac{x_0 x_2}{2N_e})] - [\frac{\partial \rho(x,p,t)(-2ux_0+2vx_1+2s_1x_0x_1)}{\partial x_0} + \frac{\partial \rho(x,p,t)(2ux_1-2vx_2-2(s_2-s_1)x_1x_2)}{\partial x_2}] \quad (16)$$

As it was shown in equation (15), the first part of (15) involves the alleles-frequency variances, the second part involves the covariances between allele frequencies, and the third part is the mean. We are using $2N_e$ in (16) instead of N_e because we are now considering diploid population. $\rho(x_0, x_2, t)$ is the probability distribution for the random variables x_0 and x_2 at time t. Solving for our probability distribution, we can see the changes in frequencies throughout our generation and for $x_0 = 0$ and $x_2 = 0$, we will be able to see the time to fixation which is the main purpose of this research.

5 Discussion and conclusions

Our study here is an extended version of what Higgs (1997) has done. In his model, Higgs assumes reversible mutation with 2 loci, each with two alleles. The two alleles are labels as in our model but in his study, Higgs assumes both the AB genotype and double mutant ab to have fitness 1, while the two single mutants Ab and aB have a reduced fitness 1-s. Let us remember that we are working in discrete generation for both our model and Higgs' model. Prior to do any modification of Higgs' model, we will first look at this model.

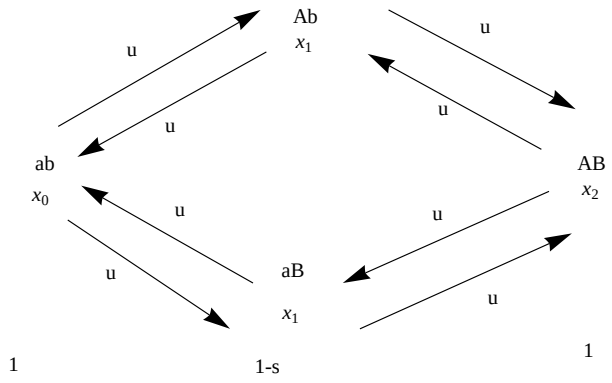

Figure 1: Higgs general model for the infinite population with reversible mutation where u=v and $s_2 = 0$

$$X_0 = (1 - 2u + 2sx_1)x_0 + 2ux_1 - r(x_0x_2 - x_1^2) \tag{17}$$

$$X_1 = (1 - 2u - s + 2sx_1)x_1 + u(x_0 + x_2) + r(x_0x_2 - x_1^2) \tag{18}$$

$$X_2 = (1 - 2u + 2sx_1)x_2 + 2ux_1 - r(x_0x_2 - x_1^2) \tag{19}$$

From this set of equations, and considering we do not have recombination, Higgs got the following set of solutions

$$x_0 = x_2 = \frac{1}{2} - \frac{u}{s} \tag{20}$$

$$x_1 = \frac{u}{s} \tag{21}$$

and the second solutions are

$$x_0 = x_2 = \frac{1}{2} - \frac{2u}{s} \tag{22}$$

$$x_1 = \frac{u}{s} \tag{23}$$

In the equations (17), (18), and (19), we have reversible mutation but the forward and backward mutation rates are equal (u=v). Also the fitnesses of the double mutant genotypes ab and AB are equal (1), and this is the same for the single mutant genotype aB and Ab (1-s). Now we will make the first step assumption by assuming $u \neq v$, but the fitnesses of the double mutant genotypes are still equal. Our model will become

$$X_0 = (1 - 2u + 2sx_1)x_0 + 2vx_1 - r(x_0x_2 - x_1^2) \tag{24}$$

$$X_1 = 2(1 - u - v - sx_0 - sx_2)x_1 + 2ux_0 + 2vx_2 + 2r(x_0x_2 - x_1^2) \tag{25}$$

$$X_2 = (1 - 2v + 2sx_1)x_2 + 2ux_1 - r(x_0x_2 - x_1^2) \tag{26}$$

and from here, we get the following solutions

$$x_0 = x_2 = \frac{-v}{s} \tag{27}$$

$$x_1 = \frac{1}{2} + \frac{u+v}{2s} \tag{28}$$

Now we will go on with our last assumption which is the main solution we are interested in. We will assume $u \neq v$ and also the double mutant genotype with different fitnesses; ab (1) and AB $(1 - s_2)$, but the single mutant genotype still have the same fitnesses $(1 - s_1)$. From here we get the equations (1), (2), and (3) with their respective solutions (10),(11), and (12).

References

[1] . Kimura, T. Ohta, 1968. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61: 763-771.

[2] .W. Strickberger (3rd edition), 2000. Evolution. United States.

[3] .H. Gillespie (2nd edition), 2004. Population genetics: a concise guide. Baltimore, Maryland.

[4] . Iwasa et al., 2005. Population genetics of tumor suppressor genes. Journal of theoretical biology 233:15-23.

[5] . Lynch, A. Abegg, 2010. The rate of establishment of Complex Adaptations. Mol. Bio. Evol. 27(6):1404-1414.

[6] . Bulmer, 1991. The selection-mutation-drift Theory of Synonymous codon usage. Genetica 129:897-907.

[7] . Lynch, 2008, Appendix 1: Diffusion Theory.

[8] . G. Higgs, 1998. Compensatory neutral mutations and the evolution of RNA. Genetica 102/103:91-101.